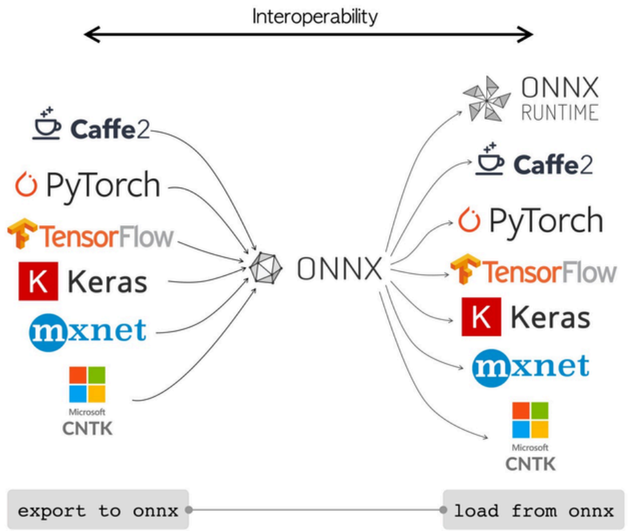

AI Models Preparation for Deployment

Ex 02

2000

To be done individually

Generated by ChatGPT 4

Exercise Overview: Model Saving, Inference, and Format Comparison with CIFAR-10

Part 1: Training and Saving the Model

Dataset Loading and Training

Load and preprocess the CIFAR-10 dataset using torchvision. Build a CNN model capable of classifying CIFAR-10 images (e.g., a simple ResNet or custom CNN).

Saving Models in Different Formats

Save the trained model in three formats:
.**pt:** Save only the model's state dictionary.
.**pth:** Save the entire model object with its architecture.
ONNX: Export and save the model to ONNX format.

Part 2: Inference on a Single Image

- Load the .pt file, rebuild the model architecture, and use it to predict the class of a single CIFAR-10 test image.
- Load the model directly from the .pth file and predict the class of a single CIFAR-10 test image.
- Use ONNX Runtime to load the ONNX model and predict the class of a single CIFAR-10 test image.

Part 3: Model Accuracy Evaluation

Write scripts to evaluate model accuracy on the CIFAR-10 test set for each saved format.
ONNX: Evaluate using ONNX Runtime.
.pt and .pth: Evaluate using PyTorch.

Part 4: Format Size and Performance Comparison

- Record and compare the file sizes of the .pt, .pth, and ONNX models.

Best Wishes